FEBS Letters 584 (2010) 3635–3643

FEBS Letters
journal homepage: www.FEBSLetters.org

From phenotype to gene: Detecting disease-specific gene functional modules via a text-based human disease phenotype network construction

Shi-Hua Zhang [a], Chao Wu [a], Xia Li [a,*], Xi Chen [a], Wei Jiang [a], Bin-Sheng Gong [a], Jiang Li [a], Yu-Qing Yan [b]

[a] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China
[b] Department of Genetics and Genome Biology, University of Toronto, Toronto, Ontario, Canada M2J 4A6

ARTICLE INFO

Article history:
Received 5 May 2010
Revised 17 July 2010
Accepted 21 July 2010
Available online 24 July 2010

Edited by Takashi Gojobori

Keywords:
Disease phenotype
Text mining
Phenotype network
Gene functional module
Genetic origin

ABSTRACT

Currently, some efforts have been devoted to the text analysis of disease phenotype data, and their results indicated that similar disease phenotypes arise from functionally related genes. These related genes work together, as a functional module, to perform a desired cellular function. We constructed a text-based human disease phenotype network and detected 82 disease-specific gene functional modules, each corresponding to a different phenotype cluster, by means of graph-based clustering and mapping from disease phenotype to gene. Since genes in such gene functional modules are functionally related and cause clinically similar diseases, they may share common genetic origin of their associated disease phenotypes. We believe the investigation may facilitate the ultimate understanding of the common pathophysiologic basis of associated diseases.

1. Introduction

In the domains of medicine and biology, phenotype data provides a valuable window for dissecting relationships between diseases and genes. Recently, many high-throughput technologies, such as systematic mutation and RNA interference, have been performed to investigate the phenotypic effect of individual genes in different species such as *Drosophila melanogaster* [1], *Saccharomyces cerevisiae* [2], *Caenorhabditis elegans* [3] and also *mammals* [4]. Today, it becomes possible to systematically analyze these productive phenotype data on a large scale in the functional genomics era.

The Online Mendelian Inheritance in Man (OMIM) [5] database, a comprehensive human disease phenotype data set, provides detailed descriptive records of different genetic diseases resulting from naturally occurring gene mutations. OMIM records are syntax-free text; there is neither a standardized vocabulary nor formal notation for the organization and representation of OMIM data. Despite such difficulties, some efforts have successfully utilized this daunting phenotype data. For instance, Freudenberg and Propping [6] clustered 878 disease phenotypes of known genetic origin from the OMIM database according to their phenotypic similarity. Their results revealed that genes leading to similar disease phenotypes have similar functional annotation. Similarly, van Driel et al. [7] compared different disease phenotypes and found that phenotype similarity correlates with various measures of gene function, such as protein sequence similarity, protein–protein interactions (PPIs), shared protein motifs and functional annotation. These investigations support a long-held assumption that genes associated with similar disease phenotypes are more likely to be functionally related. These functionally related genes serve together in a biological functional module, such as protein complex, pathway and cellular organelle, to perform a desired cellular function [8–10]. Inspired by the indication of these results, we attempted a promising work to establish the relationship between disease phenotypes and their underlying gene functional modules since it can help to uncover the molecular mechanisms of most genetic disease.

In this study, we aimed to develop a text-based phenotype network modeling method to investigate the relationship between disease phenotypes and their underlying gene functional modules. To this end, we chose 2136 phenotype records with known disease genes from the OMIM database, and constructed the human disease phenotype network. In the network, we extracted 102 phenotype clusters using a graph-based clustering algorithm. Within each phenotype cluster, we mapped disease phenotypes to genes using the disease-gene association. Thus, we created 102 gene subsets mapped to 102 corresponding phenotype clusters. Of these 102 mapped gene subsets, 82 (80.39%) showed enrichment in Gene

* Corresponding author. Address: College of Bioinformatics Science and Technology, Harbin Medical University, 194 Xuefu Road, Nangang Distinct, Harbin 150081, China. Fax: +86 451 8661 5922.
E-mail address: lixia@hrbmu.edu.cn (X. Li).

0014-5793/$36.00
doi:10.1016/j.febslet.2010.07.038

Ontology (GO) analysis [11]. Genes involved in such subsets are functionally related and represent the shared genetic origin of each of the phenotype clusters. We can call them disease-specific gene functional modules corresponding to different phenotype clusters.

2. Materials and methods

2.1. Human disease phenotype data

In the current OMIM database, there are more than 19 000 full-text records, every record corresponding to one gene or one disease phenotype. As to phenotype records, different names of disease phenotypes with the same OMIM ID were pooled into a single disease phenotype. For example, Alzheimer disease 6 and Alzheimer disease 8, with the same OMIM ID 104 300, were regard as the same disease phenotype. Finally, we collected a total of 2136 phenotype records with a unique OMIM ID for each one.

2.2. Constructing of disease phenotype feature vectors

Phenotype records contain some fields for description of genetic diseases. In this method, we considered the combination of text (TX) and clinical synopsis (CS) fields as a phenotype record. These phenotype records were automatically parsed by the MetaMap Transfer tool [12], a highly configurable program to map biomedical text to the Unified Medical Language System (UMLS) [13] Metathesaurus concepts. Thus, phenotype records could be represented by corresponding biomedical term vectors. We referred to these term vectors as phenotype feature vectors because such terms/concepts can serve as phenotypic features to characterize different genetic diseases. To ensure phenotype feature vector be more relevant to biomedical terms, some clinically irrelevant semantic types, e.g., STY (UI: T065): Educational Activity, STY (UI: T093): Health Care Related Organization, STY (UI: T066): Machine Activity and so on, were filtered out in the parsing. For the following phenotypic similarity computation, we applied the term frequency-inverse document frequency (TF-IDF) weighting scheme [14] for the refinement of these phenotype feature vectors. In this scheme, the feature weights for each phenotype were the local and global combining weights, and the augmented normalized term frequency was used as an amendment for the local weight (see Supplementary data).

2.3. Disease phenotype similarity score

To quantitatively describe the phenotypic similarity between different phenotype record P_j and P_k, we defined the similarity measure as cosine of the angle between their corresponding phenotype feature vectors using the following formula:

$$\text{Similarity}\ (P_j, P_k) = \frac{\sum_{i=1}^{N} w_{i,j} * w_{i,k}}{\sqrt{\sum_{i=1}^{N} (w_{i,j})^2} * \sqrt{\sum_{i=1}^{N} (w_{i,k})^2}}$$

where N was the sum of mapped UMLS concepts, $w_{i,j}$ and $w_{i,k}$ were the ith term' weight in phenotype record P_j and P_k, respectively.

2.4. Constructing method of the phenotype network

The construction of the phenotype network was based on the phenotypic similarity score between different disease phenotypes. In the phenotype network, the association between any two different disease phenotypes was determined when their phenotypic similarity score exceeded the significant cutoff. To achieve the cutoff, we firstly randomly shuffled the order of weighs in the corresponding two phenotype feature vectors 1000 times, and calculated the similarity scores of pairs of shuffled phenotype records. Then these similarity scores were ranked in descending order and finally, the minimal score of the top five percent (empirical P-value is 0.05) was chosen as the significant cutoff.

2.5. Detecting disease-specific gene functional modules

From the phenotype network, we used the graph-based clustering algorithm, described by Bader et al. [15], to detect densely connected subgraphs that we called phenotype clusters because nodes of a subgraph represent similar biological disease phenotypes. The applied method can give scored results, and the scoring measure is defined as density D of identified subgraphs. Density D reflected the connectivity level of a subgraph, so D was defined as the number of edges E divided by the possible maximum number of edges E_{max} of a subgraph. For each of the phenotype clusters, mapping was implemented from disease phenotypes to their associated disease genes based on the disease-gene association list in the OMIM database. Thus, we can get the corresponding gene subsets mapped to different phenotype clusters.

To perceive functional significance of mapped gene subsets in the context of the phenotype network. DAVID [16] GO-term enrichment analysis was performed for each of them. It is known that GO functional annotation system is well-organized in a hierarchy structure; the deeper the GO annotation level is, the more specific the annotated biological function is. Thus, we chose GO terms at the fifth annotation level [16], which represents specific and informative functional categories, to explore the specific functional relationship among genes in different gene subsets. The fact that some gene subsets show enrichment in GO analysis illuminated that genes in such subsets were functionally related and represented the shared genetic origin of each of the phenotype clusters. We can call them disease-specific gene functional modules corresponding to different phenotype clusters.

2.6. Disease class enrichment analysis of gene functional modules

We used the disease classification established by Goh et al. [17], who manually classified OMIM diseases into 22 disease classes according to the physiological system affected, to conduct the disease class enrichment analysis, which was implemented to investigate whether genes (molecular level) in a given gene functional module tend to have their associated disease phenotypes (phenotype level) belonging to the same disease class (i.e., specificity of gene functional module to the assigned disease class). For a gene, the disease class of the associated disease phenotype was considered as its class attribute. In our method, we randomly shuffled the class attributes of genes in a gene functional module. Therefore, the framework could be devised as follows: (i) for the corresponding phenotype cluster of a gene functional module, we randomly picked from all the disease phenotypes and built 10 000 pseudo phenotype clusters that have the same number of disease phenotypes as the real phenotype cluster, (ii) in the real phenotype cluster, the possible disease classes were decided and the number of disease phenotypes belonging to each certain disease class was counted and (iii) the P-values for every possible disease class decided in the real phenotype cluster was calculated based on the random controls.

2.7. Evaluation of gene functional modules with PPI intensity

It can be speculated that gene products (proteins) within the same gene functional module tend to interact with each other at the PPI functional level [7]. Thus, we introduced the measure of PPI intensity I_{ppi} for the evaluation of the detected gene functional

modules. The measure was defined as the fraction of actual existing PPIs among the possible maximum number of PPIs in a given gene functional module *i*. Thus, it can be formulated as follow:

$$I_{ppi} = 2N_{i\_actual}/(k_i(k_i - 1))$$

where N_{i_actual} was the actual existing PPI number between gene products in the gene functional module *i*, k_i was the number of gene products in this gene functional module which can be found having interactions with others. Note that when calculating N_{i_actual}, we considered interactions with at most one intermediate; that is, two levels of interactions: direct and indirect (the distance is 2) PPI were counted. Thus, we could get the mean I_{ppi} for all the gene functional modules. In order to test the statistical significance of the obtained mean I_{ppi}, *M* (the number of detected gene functional modules) gene sets (same size as the corresponding gene functional modules) were chosen from all the mapped genes as a random control, and eventually 10 000 such controls were built.

3. Results

3.1. The modular nature of the phenotype network

The constructed phenotype network (Fig. 1) contains 19 455 associations among 1809 disease phenotypes, with a giant connected component of 1598 (88.34%) disease phenotypes and 19 241 (98.90%) associations. In the network, disease phenotype nodes were marked with different colors based on their assigned disease classes [17]. It is clear that disease phenotypes were more likely to associate with those of the same disease class. This was consistent with the previous result of Goh et al., indicating that disease phenotypes of the same class tend to have shared genetic origin and form local functional clustering (modularity).

We used two measures of modularity: dyadicity *D* and heterophilicity *H* (see Supplementary data), which were proposed by Park and Barabasi [18] to quantify the modular properties in the phenotype network. Dyadicity is a measure of the enrichment of links between nodes sharing a common property over the number expected if the characteristics were distributed randomly on the network. Heterophilicity is a measure of the tendency of nodes to connect with other nodes with a common property. In the phenotype network, disease phenotypes of the same disease class were regarded to have the common property. Thus, we can compute the *D*s and *H*s for the 21 main disease classes (not considering the unclassified class). The phenotype network has a highly modular structure, as demonstrated by the finding that all the disease classes are dyadic ($D > 1$) and most (81%) are heterophobic ($H < 1$, Table 1), indicating that they have distinct properties in the genetic origin. However, several disease classes, such as multiple, developmental, skeletal and dermatological diseases, were heterophilic ($H > 1$), indicating that these disease classes have overlapping clinical phenotypes with other categories of diseases. This would be for multiple diseases because they arise as a consequence of the disfunctions of multiple tissues. For developmental diseases, this arises because they cause pathological changes in multiple tissues. For heterophilic diseases (skeletal and dermatological), we speculated that such heterophilic diseases may affect other tissues during the disease course, and thereby overlap with other classes of diseases.

3.2. Topological properties of the phenotype network

The phenotype network showed an obvious scale-free property [19] because it had a degree-distribution of power-law (Fig. 2A). Of the total 1809 disease phenotypes, only 76 (0.036%, act as "hubs") have more than 100 associations with other disease phenotypes,

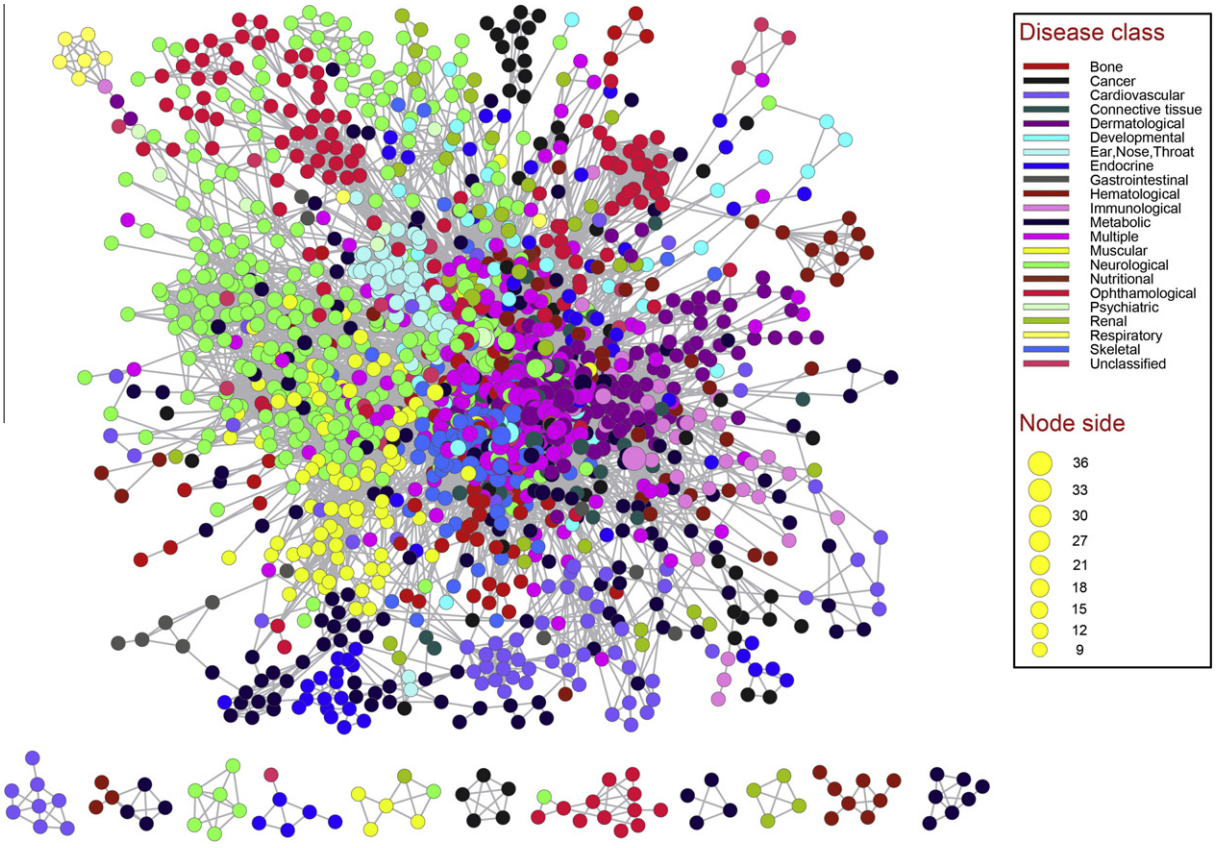

Fig. 1. Visualization of the phenotype network. In the phenotype network, the color of a disease phenotype node depends on the disease class to which it belongs and the size of it is proportional to its degree.

Table 1
Dyadicity(H) and heterophilicity (D) values of 21 disease classes.

Disease class	Disease phenotype[a]	In-class links	Out-class links	D Value	H Value
Bone	47	105	913	6.3976	0.7815
Cancer	102	650	491	8.3114	0.2004
Cardiovascular	87	251	688	4.4191	0.3261
Connective tissue	30	89	807	13.4758	0.7348
Dermatological	95	710	2446	10.4734	1.0672
Developmental	40	55	1328	4.6443	1.3301
Ear, nose, throat	43	523	797	38.1477	0.7439
Endocrine	88	339	568	5.8328	0.2663
Gastrointestinal	34	100	151	11.740	0.1772
Hematological	64	123	405	4.0185	0.2572
Immunological	67	270	850	8.0432	0.5167
Metabolic	186	323	2048	1.2365	0.4841
Multiple	156	902	4975	4.9139	1.3746
Muscular	57	139	842	5.7363	0.5980
Neurological	273	1617	4298	2.8685	0.7349
Nutritional	19	153	62	58.9318	0.1290
Ophthalmological	118	567	964	5.4100	0.3436
Psychiatric	41	209	890	16.7875	0.8702
Renal	43	58	410	4.2305	0.3827
Respiratory	23	139	240	36.1867	0.4137
Skeletal	49	262	2195	14.6740	1.8045

[a] The number of disease phenotypes belonging to the left listed disease class.

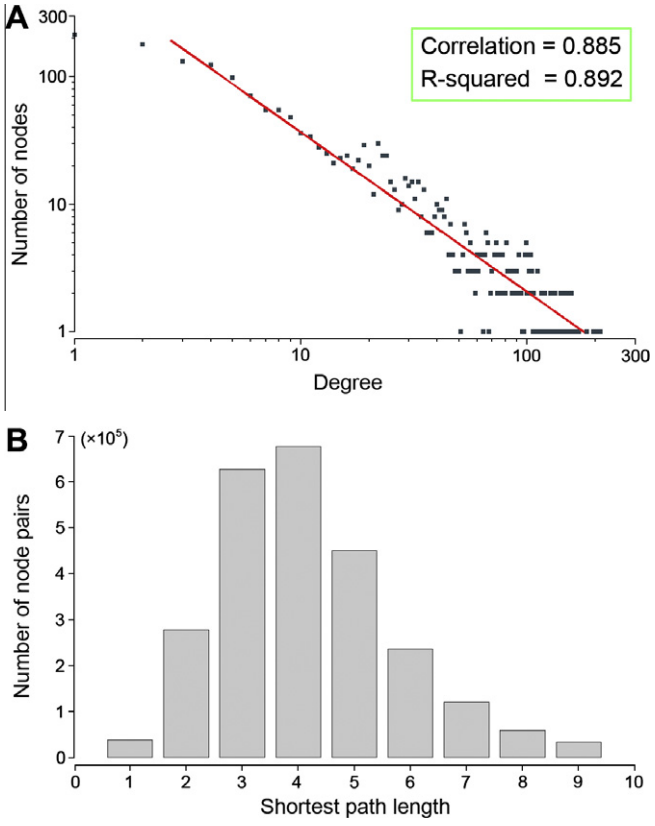

Fig. 2. Topological analysis of the phenotype network. (A) In the plot, double logarithmic coordinates are chosen. (B) The plot gives the number of node pairs with the shortest path length l ranging from 1 to 9.

suggesting that these few highly connected hubs held the network together. Interestingly, most of these disease hubs, such as Schwartz-Jampel syndrome (degree $k = 114$) and Stickler syndrome ($k = 111$), belong to the multiple disease class (Supplementary Fig. S1). The fact that the multiple disease class has more associations than those of other disease classes in the network is reasonable because they arise due to dysfunctions of multiple tissues and therefore have overlapping clinical phenotypes with different classes of diseases.

Fig. 2B shows the shortest path length distribution of the phenotype network. There are 2552 760 (77.9%) node pairs with a shortest length l of 1–9. By calculation, the network has a short mean-shortest path length of 3.5. Thus, any two nodes in the network can be connected with a path of only a few links. In addition, the average clustering coefficient $\bar{C}$ of the network is 0.24. To determine the statistical significance of the observed value of $\bar{C}$, degree-preserving random shuffling was performed for the network 10^4 times, and the result showed that $\bar{C}$ is significantly higher than random control (P-value < 10^{-3}). Together, these results indicate that the phenotype network has a small-world property [20]. The small-world property, on the other hand, revealed the close genetic relationships between different disease phenotypes, even though most are not neighbors in the phenotype network.

3.3. Disease-specific gene functional modules

We extracted 102 phenotype clusters from the phenotype network. It is to be noted that a relatively high proportion (66.67%) of phenotype clusters have 3–5 disease phenotypes (Fig. 3), indicating that most disease phenotypes tend to form small phenotype clusters; this also has specific implication in the common genetic origin. In addition, our statistical results showed that disease phenotypes in a given phenotype cluster tend to belong to the same disease class (Supplementary Fig. S2). This is consistent with the previous visual indication in the phenotype network.

To gain more insight into the shared molecular mechanism of associated human genetic diseases, mapping was implemented from disease phenotype to gene based on the disease-gene association. For 102 gene subsets mapped to corresponding 102 phenotype clusters, DAVID GO-term enrichment analysis was performed, and the results indicated that 82 (80.39%) gene subsets show enrichment (P-value $\leqslant 0.05$). Genes in such subsets (as different gene groups) may together perform distinct biological functions; that is, these genes are functionally related and represent the shared genetic origins of their associated disease phenotypes. We can call them disease-specific gene functional modules corresponding to different phenotype clusters. It is interesting that potential functions of these gene functional modules are closely associated with phenotypic traits of the corresponding phenotype

clusters. For example, phenotype cluster 62 in Table 2 shows that five disease phenotypes of it all belong to metabolic disease; this is logical because the corresponding gene functional module is enriched in a metabolism-related GO functional category: carboxylic acid metabolic process (GO:0019752).

3.4. Gene functional modules have significant disease class specificity

We referred to the disease class annotations established by Goh et al. [17] to conduct the disease class enrichment analysis. Our results showed that 68 (82%) gene functional modules were significantly enriched in only one disease class, 10 gene functional modules (12%) in 2 disease classes and only 4 gene functional modules (4.8%) in 3 or 4 disease classes (Supplementary Table S1). These four gene functional modules were relatively large and could be decomposed into several smaller modules. We prefer to regard them as different hierarchical structures consisting of smaller gene functional modules in the gene network. In this way, such hierarchical structure may perform several different biological functions, and ultimately cooperate to complete a desired cellular process. Together, these results show that the majority (94%) of detected gene functional modules have significant specificity to certain disease classes, indicating that these gene functional modules represent their shared genetic basis and that related genes in a gene functional module tend to form local functional clustering in the gene functional network.

3.5. Gene products in a gene functional module tend to interact

The PPI data were downloaded from the Human Protein Reference Database (HPRD) [21]. Of the total 82 gene functional modules, 5 were excluded because none of their members has interactions with others in the HPRD. Of the remainder, 15 gene functional modules have $I_{ppi} = 0$ or ≈ 0, and 62 (81%) have I_{ppi} greater than 0.1 (6 of the 62 gene functional modules have $I_{ppi} = 1$). The mean I_{ppi} is 0.21, which is significantly higher than that of random controls (P-value = 3 E−4). These results indicated that gene products in a gene functional module have a tendency to interact with each other, and are part of the same biological process. Gene products in such gene functional modules may participate in the same cellular pathway or molecular complex in the form of a functional module. Therefore, dysfunction of a gene functional module associated with a certain biological function will lead to several relevant diseases.

3.6. GO similarity of gene functional modules is well correlated with phenotypic similarity

For detected gene functional modules and their corresponding phenotype clusters, two functional measures can be calculated. One is the similarity of GO terms assigned to genes in a gene functional module (GO similarity) and the other is phenotypic similarity of disease phenotypes in a phenotype cluster (phenotype similarity). In our approach, a gene functional module and its corresponding phenotype cluster were regarded as the same object. Hence, we could calculate for each gene functional module the average GO similarity and the average phenotypic similarity score. These two kinds of similarity scores of the detected 82 gene functional modules were represented as two corresponding 82-dimension similarity vectors, and therefore we could get the correlation score of the GO similarity of the 82 gene functional modules with the phenotypic similarity using the correlation of these two similarity vectors.

GO similarity score was calculated using the Jiang–Conrath method [22,23] available in the R-package GOSim [24], and the correlation measure used was the commonly known Pearson correlation coefficient r. To test the statistical significance of the obtained r value, we built 1000 control groups for each of the gene functional modules and phenotype clusters from randomly picked disease phenotypes or disease genes. The r value was 0.47 (P-value < 10^{-3}) and the result indicated that GO similarity and phenotypic similarity have a relative high positive correlation. This high correlation indicates that the functional relationships among genes in a gene functional module are consistent with the phenotypic relationships of disease phenotypes in the corresponding phenotype cluster; that is, related genetic diseases share similar

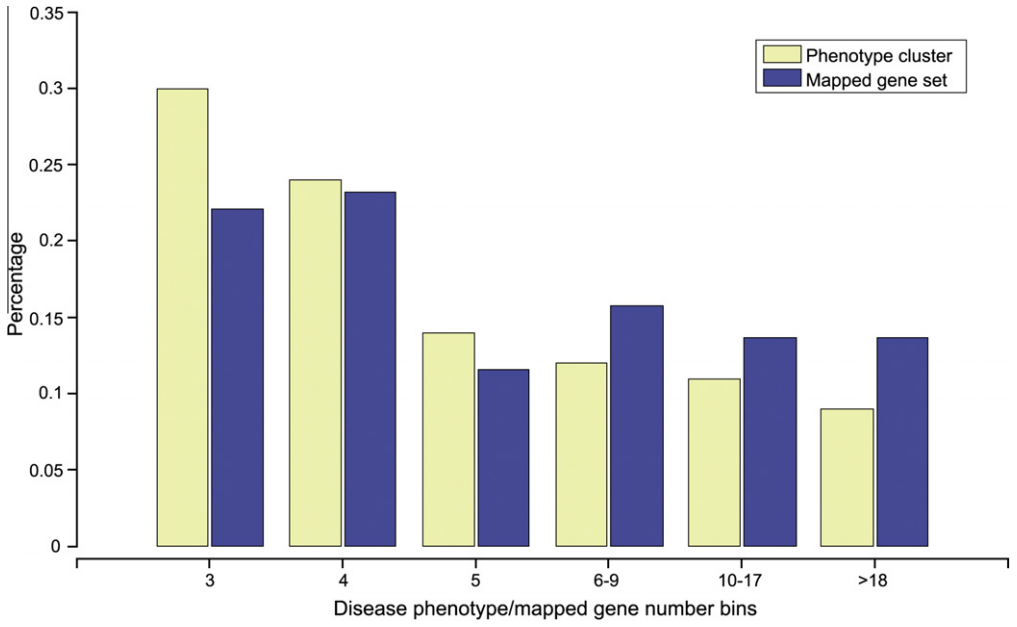

Fig. 3. Distribution of the sizes of phenotype cluster and mapped gene subset. The diagram shows the distribution of the number of phenotype clusters and their corresponding mapped gene subsets in different number bins.

Table 2
Ten typical phenotype clusters and their corresponding gene functional modules at the level of molecular function.

Disease phenotype level				Gene molecular level			
ID[a]	Illus[b]	Density	Disease class	# MG[c]	%Hits[d]	*P*-Value	Potential function
8		0.4500	Dermatological	11	30	2.7E−3	Epidermis development
19		0.4167	Ophthalmological	6	100	1.2E−9	Sensory perception of light stimulus
35		0.3333	Cardiovascular	11	36.4	4.9E−6	Heart contraction
42		0.3200	Neurological	12	27.3	9.7E−3	Second-messenger-mediated signaling
46		0.1250	Muscular	12	50	1.2E−6	Muscle development
61		0.3750	Bone	5	40	4.5E−3	Bone mineralization
62		0.2800	Metabolic	6	83.3	9.2E−6	Carboxylic acid metabolic process
70		0.3125	Endocrine	5	60	9.5E−4	Gamete generation
71		0.3125	Metabolic	5	60	1.4E−2	Lipid metabolic process
98		0.3333	Developmental	6	80	1.3E−2	Multicellular organismal development

[a] ID (identifier) is sequence number of the 102 phenotype clusters extracted from the phenotype network, this is different from that of the detected 82 gene functional modules.
[b] Illus is the abbreviation of illustration, the Illus[b] column shows structure subgraphs of the 10 typical phenotype clusters for visualization.
[c] MG is the abbreviation of mapped genes, the # MG column shows the number of mapped genes in these 10 different gene functional modules.
[d] %Hits shows the percentage of genes that are in concordance with the potential function of a gene functional module, this can facilitate a critical assessment of the obtained results.

molecular mechanism. Thus, the consideration of common therapeutic method and universal drug use should be noted in clinical therapy for associated diseases.

3.7. Gene functional modules of the same disease class have shared GO categories

In our study, we used the bi-clustering method to gain a relation profile of the obtained gene functional modules based on their over-represented GO biological process (BP) terms/categories. The profile can help to examine the shared molecular mechanism of different gene functional modules that contribute to certain diseases. Thus, we can further investigate the close relation in pathogenesis between these different diseases. Fig. 4 shows the heatmap of bi-clustering for these 82 gene functional modules (upper) and their corresponding 93 over-represented GO categories (right). The top color gradient represents $-\log P$-values of gene functional modules enriching in each GO category. It is interesting that gene functional modules with corresponding phenotype clusters belonging to the same disease class tend to form larger module groups and similar results are seen with function-related GO categories (indicative in the colored bars in Fig. 4). For example, the purple rectangle region (labeled in A) shows that gene functional module 44, 75 and 82 are enriched in two related GO categories: ectoderm development (GO:0007398) and epidermis development (GO:0008544, see enlarged region with light blue shading). These three gene functional modules represent functional entities associated with development of skin tissue. In this case, the perturbation or breakdown of these entities will produce recognizable dermatological abnormality. This is logical because the corresponding phenotype clusters of these three gene functional modules belong to the same dermatological disease class. In the view of relation profile of gene functional modules, we surmise that these three gene functional modules may functionally interact and form a hierarchical organization to perform a set of systematic functions.

Fig. 4. Bi-clustering of gene functional modules and their corresponding over-represented GO categories. On the upper, horizontal bars (aligned) of different colors are placed below the identifiers (IDs) of the 82 gene functional modules (abbreviated as GFM). The bar color depend on the disease class (indicated on top of these GFMs with a solidus mark) to which the corresponding phenotype cluster belongs.

4. Discussion

In this work, we presented a text-based framework to establish the associations between genetic diseases and gene functional modules, which may help to uncover the molecular mechanisms of genetic diseases. The framework integratively analyzed the associations of phenotype/phenotype, phenotype/gene and gene/function. Consequently, we obtained disease-spe-

cific gene functional modules which represented functional entities of associated disease phenotypes. Thus, our work may facilitate the investigation of genetic principles of human diseases from a new view. Recently, some efforts have been exerted to construct a human disease network. To reveal the relationships between different human genetic diseases, Goh et al. [17] extracted the known disease-gene association list in the OMIM database, and constructed a human disease network. In their method, two diseases are linked if they share at least one disease gene. Lee et al. [25] constructed a cell metabolism-based human disease network in which two diseases are linked if mutated enzymes associated with them catalyze adjacent metabolic reactions. The two approaches used disease genes (or enzymes) as intermediates to determine disease interactions. The present study focused on phenotype data, which can provide multiple levels of information, such as genetic factor, pathogenesis and clinical feature. Thus, text-based phenotypic similarity may be a valuable measure to evaluate the associations of different disease phenotypes. Based on the constructed phenotype network, we identified phenotype clusters, from which gene functional modules were obtained using the associations between phenotypes and genes. Since considering the phenotype information, this strategy was different from previous module mining methods [26–29], which usually used data at the molecular level, such as PPI data [30–32]. In addition, gene functional modules were further evaluated from the aspects of phenotype classification and protein function by using disease class enrichment and PPI intension analysis, respectively. Our results showed that gene function modules have a high consistence in these two aspects.

The topological properties (such as scale-free and small-world properties) display in the phenotype network are useful for detecting general interaction patterns among diverse disease phenotypes. For example, the fact that the multiple disease class tends to act as the network hubs indicates the existence of a critical disease class in the disease association map. At the level of modeled function from the phenotype network, we detected 82 disease-specific gene functional modules. As different gene groups, these gene functional modules are shown to be functionally related by the GO and PPI intensity analysis results; thus they represent the basic functional units (e.g., multi-protein complex) of biological systems and perform distinct cellular functions. As expected, gene functional modules and the corresponding phenotype clusters show a high level of agreement in functional interplay, although they are at two different biological levels. We have presented examples (see Supplementary data) of this agreement, which likely reflects the close casual relationship between genetic disease and gene functional module.

With increasing amounts of disease phenotype data available, the identification of gene functional modules may be further improved. These results provide a potentially valuable association profile of disease phenotypes and their underlying gene functional modules, which can further our understanding of the shared molecular mechanism of associated diseases. In the future, we can consider an possible applications of the association profile, for example, including the use of distinct levels of biological data, such as gene expression, proteins interactions and GO annotation, to conduct candidate gene prediction in an integrated network.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 30871394, 30370798 and 30571034), the National High Tech Development Project of China, the 863 Program (Grant No. 2007AA02Z329), the National Basic Research Program of China, the 973 Program (Grant No. 2008CB517302), the Educational Foundation of Heilongjiang Province (Grant No. 11541298) and the National Science Foundation of Heilongjiang Province (Grant Nos. JC200711, ZD200816-01, ZJG0501 and GB03C602-4).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2010.07.038.

References

[1] Kuttenkeuler, D. and Boutros, M. (2004) Genome-wide RNAi as a route to gene function in Drosophila. Brief Funct. Genomic Proteomic 3, 168–176.
[2] Giaever, G. et al. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. Nature 418, 387–391.
[3] Rual, J.F. et al. (2004) Toward improving *Caenorhabditis elegans* phenome mapping with an ORFeome-based RNAi library. Genome Res. 14, 2162–2168.
[4] Shi, Y. (2003) Mammalian RNAi for the masses. Trends Genet. 19, 9–12.
[5] Hamosh, A., Scott, A.F., Amberger, J., Bocchini, C., Valle, D. and McKusick, V.A. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 30, 52–55.
[6] Freudenberg, J. and Propping, P. (2002) A similarity-based method for genome-wide prediction of disease-relevant human genes. Bioinformatics 18 (Suppl. 2), S110–S115.
[7] van Driel, M.A., Bruggeman, J., Vriend, G., Brunner, H.G. and Leunissen, J.A. (2006) A text-mining analysis of the human phenome. Eur. J. Hum. Genet. 14, 535–542.
[8] Badano, J.L. and Katsanis, N. (2002) Beyond Mendel: an evolving view of human genetic disease transmission. Nat. Rev. Genet. 3, 779–789.
[9] Gandhi, T.K. et al. (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat. Genet. 38, 285–293.
[10] Kann, M.G. (2007) Protein interactions and disease: computational approaches to uncover the etiology of diseases. Brief Bioinform. 8, 333–346.
[11] Blake, J.A. and Harris, M.A. (2002) The Gene Ontology (GO) project: structured vocabularies for molecular biology and their application to genome and expression analysis. Curr. Protocol Bioinformatics, Chapter 7, Unit 72.
[12] Aronson, A.R. (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc. AMIA Symp., 17–21.
[13] Bodenreider, O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 32, D267–D270.
[14] Wilbur, W.J. and Yang, Y. (1996) An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. Comput. Biol. Med. 26, 209–222.
[15] Bader, G.D. et al. (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4, 2.
[16] Dennis Jr., G., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C. and Lempicki, R.A. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 4, P3.
[17] Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M. and Barabasi, A.L. (2007) The human disease network. Proc. Natl. Acad. Sci. USA 104, 8685–8690.
[18] Park, J. and Barabasi, A.L. (2007) Distribution of node characteristics in complex networks. Proc. Natl. Acad. Sci. USA 104, 17916–17920.
[19] Barabasi, A.L. and Bonabeau, E. (2003) Scale-free networks. Sci. Am. 288, 60–69.
[20] Gong, Y. and Zhang, Z. (2009) Global robustness and identifiability of random, scale-free, and small-world networks. Ann. NY. Acad. Sci. 1158, 82–92.
[21] Peri, S. et al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res. 13, 2363–2371.
[22] Jiang, J.J. and Conrath, D.W. (1997) Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. ROCLING X.9008.
[23] Guo, X., Liu, R., Shriver, C.D., Hu, H. and Liebman, M.N. (2006) Assessing semantic similarity measures for the characterization of human regulatory pathways. Bioinformatics 22, 967–973.
[24] Frohlich, H., Speer, N., Poustka, A. and Beissbarth, T. (2007) GOSim – an R-package for computation of information theoretic GO similarities between terms and gene products. BMC Bioinformatics 8, 166.
[25] Lee, D.S., Park, J., Kay, K.A., Christakis, N.A., Oltvai, Z.N. and Barabasi, A.L. (2008) The implications of human metabolic network topology for disease comorbidity. Proc. Natl. Acad. Sci. USA 105, 9880–9885.
[26] Chen, J. and Yuan, B. (2006) Detecting functional modules in the yeast protein–protein interaction network. Bioinformatics 22, 2283–2290.
[27] Hwang, W., Cho, Y.R., Zhang, A. and Ramanathan, M. (2006) A novel functional module detection algorithm for protein–protein interaction networks. Algorithms Mol. Biol. 1, 24.
[28] Luo, F., Yang, Y., Chen, C.F., Chang, R., Zhou, J. and Scheuermann, R.H. (2007) Modular organization of protein interaction networks. Bioinformatics 23, 207–214.
[29] Ulitsky, I. and Shamir, R. (2007) Identification of functional modules using network topology and high-throughput data. BMC Syst. Biol. 1, 8.

[30] Brown, K.R. and Jurisica, I. (2005) Online predicted human interaction database. Bioinformatics 21, 2076–2082.
[31] Rual, J.F. et al. (2005) Towards a proteome-scale map of the human protein–protein interaction network. Nature 437, 1173–1178.
[32] Stelzl, U. et al. (2005) A human protein–protein interaction network: a resource for annotating the proteome. Cell 122, 957–968.